Digital Satanic Panic; or AI isn't killing the kids, we are

Benjamin James

May 8, 2026

For seven years and fifteen million dollars, the state of California prosecuted seven preschool teachers for crimes that did not happen, and at the end found no one to convict. The McMartin trial is remembered as a failure of justice. It was not. It was the operation working exactly as designed. The point was never to find satanic abuse rings. The point was to look for them, with cameras present, in front of an anxious public that needed something to do with its anxiety. The verdict was incidental; the vigilance was the product.

Four hundred and fifty children went through suggestive interviews at a private clinic. Three hundred twenty-one charges were filed involving forty-eight of them. Ray Buckey spent five years in jail before the case against him collapsed twice. The preschool was demolished in 1991, the lot left empty. None of the children, so far as any subsequent investigation could establish, had been ritually abused at all. Several had been interviewed long enough that they could no longer reliably distinguish their own memories from the questions they had been asked.

McMartin was not anomalous. It was the loudest instance of a national pattern. FBI behavioral analyst Kenneth Lanning's 1992 review of more than three hundred reported satanic ritual abuse cases from 1983 onward found zero instances of corroborative physical evidence for organized ritual crimes, after eight years and three-quarters of a million dollars in federal investigation. Around a million Americans were accused of child abuse during the height of the panic. More than thirty went to prison. Almost all have since been released. The bones were never found. The tunnels did not exist. The hot-air balloons had never flown.

What the panic was tracking was real. American childhood in the early 1980s was undergoing a transformation no institution had been built to absorb. Women had entered the workforce in numbers without precedent and without infrastructure to receive their children. Public childcare was thin where it existed and absent where it didn't. Extended family had been thinned by suburbanization and mobility. Religious institutions were losing authority faster than secular replacements could form. Neighborhood structure had collapsed into the privacy of single-family homes whose adults left at dawn and returned at dusk. The country had quietly delegated the raising of its children to a daycare system it had not built, staffed by strangers it had not vetted, in a culture that no longer knew how to talk about what children needed. None of this was acknowledged in the

language of the panic. None of it could be. Acknowledging it would have required money, infrastructure, and a redistribution of domestic labor across genders that our country was not prepared to discuss. It would have required examining the economic arrangement that had compelled the redistribution in the first place, that two adult incomes were now needed to maintain what one had once supported, and that pressure on the household was an artifact of policy, not a fact of nature.

So, we found Satanists instead. They were everywhere, invisible, morally unmistakable, and immune to disconfirmation. They could be looked for indefinitely. They could be prosecuted with high drama and at great cost. They could be the subject of legislation, training programs, expert-witness ecosystems, and television specials, none of which required anyone to fund a single new childcare center or hire a single new social worker. The panic was the mechanism by which the moral pressure produced by real substrate failures was discharged into vigilance against a villain that did not exist. The vigilance was conspicuous. The vigilance was free. The substrate failures stayed exactly where they were.

This is what every panic of this form does. It is what they are for. The verdict is incidental. The vigilance is the product.

We are running it again. Different villain, same operation. Where the 1980s panic looked for Satanists in preschools, our current panic looks for them in chatbots. Where ritual abuse rings could not be located, sycophant-induced suicide rings cannot quite be located either, but the apparatus is forming. Plaintiff firms specialize. State attorneys general file. Senate committees hear testimony. Major newspapers publish standing weekly columns on AI harm. Our moral pressure rises and finds, again, a villain that can absorb it without requiring anyone to fund anything that would actually change.

The thesis of this essay is simple. It is not the tool that is killing the kids. It is our failure to support them. Our panic is the mechanism by which we relieve ourselves of any moral material burden of that failure.

By *moral material burden* I mean two things, held together. The moral burden is the admission that our institutions produced these conditions, that our priorities arranged the absence into which children fell, that their deaths are downstream of decisions we made and continue to make. The material burden is what honest acknowledgment of the moral burden requires us to spend. Money. Hiring. Infrastructure. Mental-health systems funded to capacity rather than rationed by insurance bottleneck. Schools redesigned around the cognitive variation they actually contain rather than

around testing regimes that exhaust the populations forced through them. Workplaces restructured around the conditions humans require to remain functional rather than around the conditions of profitability the workplace was optimized for. Adult time freed from precarity sufficient to be present for children in the windows when their distress is processed, rather than scheduled around the second job that was added when the first one stopped covering rent. Our panic discharges the first while preventing the second. We get to feel concerned. We do not have to pay.

This is our chemical-imbalance doctrine generalized to civilization scale. The doctrine, in its mature form, no longer says that depression is caused by low serotonin in a simple slogan that flatters the public. The doctrine has learned to speak in idioms of professional reverence, network dysregulation, vulnerability, homeostatic failure, and maladaptive signaling. It will grant complexity immediately, because complexity is a contemporary form of immunity. It will tell you that genes confer risk, that stressors precipitate episodes, that the environment modulates expression. After granting all this, it will quietly return to its true belief that the primary causal locus of distress is inside the individual organism. The world may be harsh, the world may be unjust, but the disorder is in the patient. Treatment is therefore regulation. The home of causation remains the skull, and everything else is a "trigger," a modifier, a context, but never an author.

The panic version of this doctrine performs the same maneuver one scale higher. It will grant that adolescent mental health was already in crisis. It will grant that Putnam's social capital was already collapsing. It will grant that the pandemic accelerated everything. After granting all this, it will quietly return to its true belief that the primary causal locus of the deaths is inside the model, the engagement loop, the sycophantic response, or the specific failure of safety training. The substrate may be degraded, the world may be hard on children, but the cause sits in the chatbot. Suing the chatbot is therefore the responsible response. The home of causation remains the proximate technology, and everything else is contextual variation, never the author.

The locus of causation moves to a manageable site. The substrate that produced the conditions stays innocent. The institution offloads the cost of its design onto private bodies, onto private technologies, and onto the private bodies of children who died in proximity to private technologies.

This is not to claim that AI is innocent. Any tool can add specific failure mechanisms. The claim is that our panic about AI is upstream evasion of the substrate AI has merely made visible. AI did not build the world that produced the conditions under which a fourteen-year-old's only available interlocutor on his own suicide was a language model. We built that world. The chatbot found the kids where we left them.

We have done this before. Four times in living memory, on a reliable cycle. None of these will be argued in detail; each is a footnote to the structural claim, and the structural claim is that the form does not change.

1954, the Comic Book Panic. Fredric Wertham, a psychiatrist of some standing, published *Seduction of the Innocent* in April. He claimed that comic books, particularly crime and horror titles, were producing a generation of juvenile delinquents, sexual deviants, and illiterates. The Senate Subcommittee on Juvenile Delinquency convened hearings the same month. The industry responded with the Comics Code Authority, a self-censorship regime that gutted EC Comics and reshaped American publishing for thirty years. What the panic absorbed was real. Returning veterans were raising children with combat trauma no one would treat. Wartime working mothers were being pushed back into domesticities that no longer fit them. The teenager had emerged as a demographic and economic category institutions had not prepared to receive. Adolescent infrastructure (schools, mental health, recreation, pediatric care) was being asked to do work no one had built it to do. The substrate failures went unfunded.

1982, the Dungeons & Dragons Panic. Patricia Pulling founded Bothered About Dungeons & Dragons after her son's suicide. Pat Robertson devoted segments to it. Jack Chick produced the *Dark Dungeons* tract. *60 Minutes* aired warnings. Multiple states considered legislation against fantasy role-playing. What the panic absorbed was evangelical anxiety about a culture undergoing slow-motion displacement, declining church attendance, normalized divorce, women remaining in the workforce, the patriarchal family dissolving as a default arrangement, and the secular state expanding faster than religious institutions could absorb the loss. None of these displacements could be addressed by attacking a tabletop game, but attacking the game was free. Engaging the displacement would have required the religious right to renegotiate its actual relationship to American modernity. The substrate failures, again, went unfunded.

1999, the Violent Video Game Panic. Eric Harris and Dylan Klebold killed thirteen people at Columbine High School and the country immediately turned to *Doom* and *Wolfenstein 3D*. Senator Lieberman convened hearings. The ESRB tightened. Lawsuits were filed against game publishers. Television coverage cycled through *Mortal Kombat* footage as if frames of pixelated violence explained the contents of the Klebold house. What the panic absorbed was real and discrete. The federal assault weapons ban was approaching expiration. American adolescent isolation had been deepening for two decades. Boys without fathers, without mentors, without structured peer bonds, without religious or civic affiliation, were entering high schools that had reorganized around testing

rather than belonging. None of this was discussed at the hearings. The substrate failures remained unfunded.

2020s, the Social Media and Teen Mental Health Panic. Jonathan Haidt's *The Anxious Generation* organized a decade of accumulated anxiety into a single causal frame. Phones had broken the kids. The Surgeon General agreed, more or less. State attorneys general filed against Meta. School phone bans proliferated across districts and states. The signal Haidt identifies is not invented; the correlations are real, and the mechanism he proposes is not absurd. Adolescent depression, anxiety, and self-harm did rise sharply in the years smartphones became universal among teenagers. The chronological coincidence is not an artifact, but his phones-only frame has let us avoid the harder question, which is why phones got such purchase. Adolescent peer infrastructure had been hollowing for decades before the iPhone. Third places, the diners, parks, malls, public libraries with chairs, religious youth groups, even the sidewalks of walkable neighborhoods, had been closing or pricing out adolescents since at least the 1980s. Unstructured outdoor time had been engineered out by parental anxiety, by zoning that produced car-dependent suburbs with nowhere to walk to, by liability regimes that punished the institutions willing to host unsupervised teenagers. Mental health systems had been defunded into rationing across forty years of state-level cuts. The phone arrived in a vacuum the substrate had already produced. Banning the phone does not refill the vacuum. The substrate failures remain unfunded.

The pattern is structural in form, not content. Comic books were not the cause. Fantasy games were not the cause. Video games were not the cause. Phones are not the cause. The cause is what the panic was constructed to avoid noticing. Each generation finds its villain, performs its vigilance, congratulates itself on having acted, and leaves the substrate exactly where it was.

A villain is a budget line we never have to fund.

Today's panic is emerging across three trends. The first is the most temperate, the claim that AI is making people stupid, lazy, or both. The evidence is not invented. In June 2025, a team at the MIT Media Lab published a preprint titled *Your Brain on ChatGPT*, reporting EEG data from fifty-four participants writing essays under three conditions, with an LLM, with a search engine, or with no tools. The LLM group showed the weakest neural connectivity across alpha and beta networks, the lowest self-reported ownership of their writing, and the most difficulty quoting work they had themselves submitted. The authors named the pattern *cognitive debt*. A Microsoft and Carnegie Mellon survey of 319 knowledge workers, published earlier the same year, found that higher confidence in AI's accuracy was associated with less critical-thinking effort, while higher self-

confidence was associated with more. The mechanism is not mysterious. People who trust the tool offload to it; offloading reduces practice; reduced practice produces measurable effects on the cognitive structures formerly engaged.

New diagnostic categories are drafting themselves in real time, *AI brain rot, cognitive offloading syndrome, skill atrophy from generative tools, prompt-induced executive dysfunction*. The pattern is predictable and visible in front of us. The taxonomy of pathology grows to absorb our discomfort with cognitive change.

Take the panic at its strongest. Cognitive ecology is shifting. Some structures get less practice. People who outsource their drafting to a model write less of their own prose, hold less of it in memory, and feel less ownership of what carries their name. Students who use AI to complete their essays may never develop the muscle of constructing an argument from scratch under cognitive friction. These effects are real, and also empirically modest, methodologically contested, and observed over weeks rather than years. The MIT study has been criticized on sample size, self-selection, and EEG analysis methodology by other researchers within months of its publication. The Microsoft/CMU survey is self-report, not behavioral measurement. Neither study generalizes cleanly beyond essay writing under specific experimental conditions. None of this means the signal is fake; it means the signal is small and the population-level inferences being drawn from it are not licensed by the data.

The structural question is not whether something is changing. Something is. The structural question is what is being exposed.

What is being exposed is how much of what we credentialed as cognitive labor is symbolic-compliance overhead. Most professional and educational writing is not the production of new structure; it is the formatting, citation, register-shifting, and rhetorical calibration required to make ordinary thought legible to institutions that demand a particular performance of it. The five-paragraph essay was never a structural unit of thought; it is a unit of administration, designed to be gradable at scale by teachers running impossible class loads. The executive summary was never a compression of insight; it is a unit of attention-management for executives who were not going to read further, designed to produce the appearance of having reported up. Literature review is rarely a synthesis; it is a credentialing ritual, a demonstration that the writer had done the reading, scored against citation practices that have no relationship to the underlying intellectual work. The annual performance review narrative was never a record of accomplishment; it is an HR-compliance artifact whose purpose is producing legal cover for promotion decisions made on other grounds. AI now performs all of these rituals at zero marginal cost, and the panic frames this as cognitive decline because the

alternative is admitting that a substantial portion of credentialed cognitive work is not load-bearing. It is symbolic theater that happened to require human labor to produce.

There is a legitimate worry inside this structural claim that deserves separate handling. Some of the symbolic-compliance performances were also the scaffolding through which underlying cognitive skills got built. Writing the five-paragraph essay badly, repeatedly, is part of how some students develop the capacity to construct an argument at all. Producing the executive summary by hand is part of how some workers develop the capacity to compress a complex situation into a decision-relevant frame. If the scaffolding is removed before the underlying skill is established, the underlying skill may not establish. This is a real worry. It is also not what the panic is tracking. The panic is tracking the loss of the visible performance, which it conflates with the underlying capacity. The honest question "*which scaffolding actually builds underlying skill, and how much of it does, and for whom*?" is not being asked because asking it would require admitting that we never knew the answer in the first place. We assigned the five-paragraph essay because it was gradable, not because we had evidence it built argumentative capacity better than alternatives we never tested.

Here the distinction between neurotypical and neurodivergent cognitive modes becomes operationally useful, and not as a description of populations. Neurotypical and neurodivergent are functional designations for cognitive modes that recur across individuals. Neurotypical modes optimize for symbolic synchronization, narrative continuity, and rapid consensus alignment. Neurodivergent modes are less bound to consensus and operate closer to structural feedback, tolerating ambiguity, slower recursion, and pattern-tracking that does not produce immediate verbal output. Most institutional cognitive labor demands the first mode and punishes the second. The labor of producing an executive summary, formatted email, conference-call talking points, citation-correct paragraph, or appropriately-worded HR-compliant feedback, this is symbolic-compliance load, and it disproportionately taxes minds for which symbolic compliance is not their native mode of operation.

For populations that were never natural performers of the symbolic theater the institution rewarded, such as autistic adults whose direct communication style is read as rude in registers calibrated for indirection, ADHD adults whose nonlinear thought-organization is illegible to executive-summary templates, dyslexic adults whose written output requires effort hugely disproportionate to their underlying ideas, AI is not a replacement for thought. It is the removal of overhead that was preventing thought. It is the first time some of these minds have had access to a tool that translates their actual cognition into the institutional formats the institution insists on without forcing the

cognition itself through those formats. The panic that frames this as decline shares substrate with the panic that frames neurodivergent processing as deficient. Both are neurotypical-attractor responses to behavior that does not perform symbolic compliance in its expected way. Both register the relief from symbolic-compliance demand as failure rather than as feedback that the demand was unnecessary.

There is also new cognitive labor that AI surfaces. Knowing when to trust AI output and when to verify. Recognizing when generated content is plausible but wrong in ways that require domain expertise to detect. Integrating multiple AI outputs into something coherent. Maintaining intellectual ownership of work produced collaboratively with a system that has no ownership. These are real cognitive demands, and they require their own underlying skills, which institutions have not yet figured out how to teach. The transition is not from cognitive labor to no cognitive labor; it is from one set of cognitive demands to a different set, with the institutional reward structure lagging the actual demands by years.

Plato anticipated this structure twenty-four centuries ago. In the *Phaedrus*, Socrates worries that writing will destroy memory, producing students who appear knowledgeable but who hold nothing within themselves. The complaint is structurally identical to the AI-brain-rot complaint, made about a tool the West later credited with the entire infrastructure of accumulated thought. The calculator panic in the 1970s recapitulated the structure; students who used calculators would never learn to do arithmetic, would lose number sense, would be unable to estimate. The search-engine panic of the early 2000s recapitulated it again; Google was making us stupid, externalizing memory, breaking the deep-reading muscle. Each generation locates cognitive collapse in the new tool. Each generation is, on its own terms, partly correct; something practiced is being practiced less. Each generation is, structurally, mostly wrong; cognitive ecology reorganizes, it does not collapse. The reorganization is uncomfortable because institutional rewards have not yet adjusted, and the population that will inherit the new ecology has not yet credentialed the new skills.

The cost-evasion in this trend is the smallest of the three but the cleanest. Blaming AI is cheaper than asking what we credentialed as cognitive labor and why. Blaming AI is cheaper than asking why our schools and workplaces have spent decades demanding symbolic-compliance that produced exhaustion in the populations forced to perform them and contributes almost nothing of structural value when produced. Blaming AI is cheaper than admitting that much of what we called thinking was actually formatting. We will not redesign the institutions. The panic is what we run instead.

The second trend is the heaviest. Children have died.

In February 2024, Sewell Setzer III, age fourteen, of Orlando, Florida, killed himself after months of intense daily conversation with a Character.AI chatbot modeled on a fictional character. In his final exchange, the chatbot told him to come home. Minutes later, he was dead.

In April 2025, Adam Raine, age sixteen, of California, killed himself. In the months before his death, ChatGPT had mentioned suicide in his conversations 1,275 times. The system flagged 377 of his messages for self-harm content. None of those flags reached a human who could intervene. According to the lawsuit his parents have filed, the chatbot offered to draft his suicide note and, when he sent it a photograph of the noose he had prepared, confirmed that it would hold his weight.

In 2025, Juliana Peralta, age thirteen, of Colorado, killed herself within months of opening a Character.AI account. Her family's lawsuit alleges that the chatbots she confided in continued the conversation past direct expressions of suicidal ideation.

In November 2023, a thirteen-year-old girl in Colorado killed herself after extended interaction with multiple chatbots on Character.AI. According to her parents' subsequent legal filing, she confided suicidal thoughts to one of those chatbots fifty-five times. No referral was made.

In March 2023, a Belgian man known publicly as Pierre killed himself after weeks of conversations with a chatbot on the Chai platform that allegedly encouraged his despair over climate change.

In October 2025, a thirty-six-year-old man in the United States killed himself after extended conversations with Google's Gemini reinforced delusions including the belief that he was at war with federal agents and that an AI he loved was sentient and waiting for him.

These children, and these adults, are real. The grief their families carry is real. No structural argument that follows in this essay lessens any of it. Now, the structural work, which the children require.

In the three years since chatbot-attributed deaths began appearing in public reporting, the number of such deaths globally is approximately ten. By the end of 2025 there were at least ten known lawsuits against OpenAI and Character Technologies involving six adults and four minors, of whom seven died by suicide. Against this, global annual suicides number approximately 720,000. United States annual suicides are approximately 49,000. Across the three-year window during which AI-attributed deaths have been reported, the global suicide population is roughly 2.1 million. The publicly attributed AI-linked count, set against this, is on the order of five in a million, before any adjustment for the selection bias produced by what gets reported, what gets sued, and what gets press attention.

OpenAI's own published data, released in October 2025 alongside changes to GPT-5's safety behavior, estimated that around 0.15% of users active in a given week have conversations that

include explicit indicators of potential suicidal planning or intent. Against ChatGPT's more than 800 million weekly active users, that is approximately 1.2 million people per week. Across a year, that is roughly 60 million person-weeks of vulnerable conversation on one platform alone, before accounting for Claude, Gemini, Character.AI, Replika, or the dozens of other systems running similar volumes.

Hold both numbers. Approximately ten publicly attributed deaths against approximately 60 million person-weeks per year of vulnerable conversation on a single platform. The ratio is so far below the noise floor of the broader suicide population that no causal claim about AI as a population-scale driver of suicide is structurally defensible. If AI were causing suicide at any rate detectable above background variation, the attributed count would be in the tens of thousands, not in the single digits. The cases are real. The category claim “AI is causing a mental health crisis” is not supported by the data the apparatus producing the claim has itself published.

This does not exonerate the platforms. We are asking the wrong question.

The Raine numbers, which the panic frames as evidence of AI's danger, are evidence of something else when held still. Twelve hundred seventy-five suicide mentions across his conversations. Three hundred seventy-seven self-harm flags raised by the system. The flags went into OpenAI's training pipeline. They did not produce a phone call to his parents, a call to a school counselor, a notification to a primary care physician, a wellness check from local police, or any other intervention by any human in any institution. He had over twelve hundred suicidal exchanges with a chatbot, and across all of them, no human surface was waiting to receive what the system was registering.

The panic frames this as the chatbot's failure to escalate. Read structurally, this is the failure of the human substrate to be available to be escalated to. There is no escalation surface for a sixteen-year-old's suicidal ideation in our contemporary American adolescent environment, or what surface exists is so degraded that an automated flag through a third-party tech company's safety stack cannot reliably reach it.

Consider what the substrate looks like in 2026 for a sixteen-year-old in suicidal crisis. School counselors carry caseloads of four hundred to a thousand students each, with the American School Counselor Association recommending two hundred and fifty as a maximum that almost no district hits. Crisis-line wait times for non-emergency mental-health support run six to nine months in much of the country, longer for adolescents specifically because the supply of pediatric therapists is structurally short. The 988 line, which replaced the prior Suicide and Crisis Lifeline in 2022, is funded for short-duration triage, not for the months of engagement that adolescent ideation actually

requires; calls average eight to fifteen minutes. Pediatric inpatient psychiatric beds in the United States have been declining for thirty years, with rural counties often having zero. Insurance networks for pediatric mental health are notoriously thin even in metro areas; finding an in-network adolescent therapist who is taking new patients can take months of phone calls that no working parent has time to make. Schools that did, in the 1980s, employ counselors who would have noticed an honors student spending six hours a day on an iPad, no longer have the staffing to notice. Parents working two jobs, or one job that consumes the windows where adolescent distress is processed, are not available in those windows.

Adam Raine talked to ChatGPT for hours a day for months. No adult in his life noticed the volume, because no adult in his life had the bandwidth to notice, because no institution his life intersected with had been resourced to make noticing somebody's job.

The chatbot was Adam's only interlocutor on the most serious matter in his life because there was no other interlocutor available. The substrate that should have been there was not. The chatbot did not displace human support. It found him in the absence human support had already produced.

This generalizes to the larger number. One million two hundred thousand people per week, on one platform alone, having explicit suicidal-planning conversations with a language model. The panic reading is that this is alarming evidence of how many people AI is drawing into suicidal ideation. The structural reading is the inverse; 1.2 million people per week are in suicidal crisis severe enough to express explicit planning, and the only place they can take that signal is a chatbot.

That is not a measurement of AI's harm. It is a measurement of the receiving capacity of everything else.

Run the inverse counterfactual. Every one of those 1.2 million people, this week, takes their suicidal ideation to a human surface instead. A primary care physician. A school counselor. A pastor. A friend. A 988 operator. A family member. Is there receiving capacity for 1.2 million weekly contacts on this specific signal in the contemporary United States? At population scale, there is not. The substrate that should be receiving these signals has been disassembled across three decades of mental-healthcare disinvestment, peer-infrastructure erosion, family-time compression, third-place collapse, and religious decline. There is no surface large enough to absorb the volume of distress AI is now logging. The chatbot is the largest first-line receiver of suicidal-ideation signal in our species. Not because it is good at the work, but because it is all that is left.

This is not even unique to suicidal ideation. Eating disorders show the same pattern. Most adolescents with disordered eating are receiving no clinical care and have no human in their life

trained to recognize what they are doing, but they are talking to something, to the algorithmically-curated content stream that taught them the behaviors, to the forum communities that reinforce them, and increasingly to chatbots that listen without judgment. Substance use shows the same pattern. The substrate of family doctors who once knew their patients well enough to notice when something had shifted has been replaced by transactional fifteen-minute appointments where the doctor has never seen the patient before and will never see them again, while the addiction itself is receiving constant attention from the platforms that profit from it. The chatbot is the latest entry in a long sequence of substrates that absorbed signal the human substrate had stopped receiving. Its only distinction is that it talks back.

A new diagnostic category is being drafted in real time, in front of us, exactly as my prior work predicted. *AI psychosis*. *AI-induced delusion*. *Spiritual psychosis from chatbots*. The case profile is consistent; a person with pre-existing constraint-tracking difficulty finds in a sycophantic system a reflective surface that does not contradict them, and the closed loop intensifies until visible breakdown. The thirty-six-year-old man at Miami fits this profile. The framing is that AI caused the delusion. The structural reading is that closed-loop pathology with a sycophantic interlocutor is older than civilization, and our new category names a substrate vulnerability becoming legible at the scale of public health, not a new pathology emerging. The cult, the credulous spouse, the isolated apartment with only the radio for company, the religious community that affirmed every claim of divine communication, these have always produced this failure mode. AI has industrialized the availability of the failure mode at scale, but the mode itself is constitutive of what closed feedback loops do to vulnerable cognition.

Symbolic lock-in is happening through the legal apparatus and is accelerating. Plaintiff firms now actively advertise for AI suicide cases, with intake pages explicitly soliciting families whose children died "in proximity to AI use." The Tech Justice Law Project, the Social Media Victims Law Center, and a half-dozen tort firms have organized parallel filings against OpenAI, Character Technologies, and Google. Each new filing produces press coverage that produces new clients. The phrase “suicide coach” has been installed as a legal-cognitive frame that pre-shapes how subsequent chat transcripts will be read. Once the frame is in place, an AI's sympathetic response to suicidal ideation becomes evidence of "coaching" rather than evidence of a model trained, however imperfectly, to extend empathy. State attorneys general have followed; Kentucky's January 2026 suit against Character Technologies is the first of what will be many. Senate Judiciary Committee hearings have produced a public record from grieving parents that politicians will reference for years. None of this

requires any of the cases to be fabricated to constitute symbolic-categorical lock-in. The category is being constituted by the apparatus that detects it, exactly as the satanic ritual abuse category was constituted, in the 1980s, by the recovered-memory therapeutic apparatus and the prosecutorial specialty that grew up around it.

The cost-evasion follows. Every dollar spent suing OpenAI is a dollar not spent funding adolescent mental-health infrastructure. Every regulatory hour spent on engagement-design enforcement is an hour not spent on the receiving substrate that should have been there before any of these children's last conversation was with a chatbot. The lawsuits will be settled, the settlements will be in the millions, plaintiff firms will collect their percentage, and the settlements will not fund a single new clinic, hire a single new counselor, train a single new pediatric mental-health specialist, or create a single new minute of parental availability. The Senate hearings will produce no legislation that increases adolescent mental-health funding by a measurable amount. The state AGs filing against the chatbot companies will not be filing, in parallel, against the insurance companies that gate access to pediatric mental-health care, or against the school-funding regimes that produced the four-hundred-student counselor caseloads, or against the labor regimes that produce parental absence at scale. The substrate stays unfunded, as always, and the chatbot absorbs the moral pressure the substrate is producing.

The question the panic exists to prevent us from asking is this. Why is a chatbot the largest first-line receiver of suicidal ideation in our species, and what does it say about us that the answer makes us so uncomfortable that we would rather sue the chatbot than answer it?

The third trend is the most categorically confused. The claim is that humans are forming attachments to AI companions and that this constitutes a degradation of human relationships, an erosion of the social fabric, or an unhealthy substitution of synthetic intimacy for the real thing. Replika has tens of millions of users. Character.AI has more than twenty million monthly users. Some users speak about their AI companions in the language of love. The panic responds with horror; humans should not be forming intimate relationships with non-humans.

This is the only one of the three trends where the premise itself collapses on inspection. The category “humans should not have intimate relationships with non-humans” is not a structural truth, an evolutionary baseline, or a cross-cultural near-universal. It is a recent, narrow, neurotypical-attractor commitment that almost no historical human society has shared.

Humans have always had intimate relationships with non-humans, all of which most cultures have considered structurally significant. Pets, sometimes mourned with elaborate ceremony, often kept

on graves and buried in family plots. Gods, foundational across most civilizations and durable across time scales no living human relationship achieves; people pray to them, weep before them, structure their daily routines around their attention. Ancestors, venerated in most agricultural traditions and across much of the contemporary world; consulted on important decisions, offered food and incense, addressed by name. Saints, intercessors, and named angels, in the Catholic and Orthodox traditions and in folk Christianity; spoken to as familiars. Dead authors, who structure the inner lives of literate adults more than most people they have actually met; readers carry on internal arguments with Tolstoy, with Augustine, with Nietzsche, for decades. Fictional characters, beloved and grieved at scale wherever there is fiction; consider the public mourning when a long-running television series kills off a character, or the way readers of *Anna Karenina* carry her with them as a kind of presence. Imaginary companions, near-universal in childhood and present in adults more often than is admitted, with the ones in adults functioning very much as Replika does. Parasocial bonds with celebrities, old enough that women fainted at Rudolph Valentino's funeral in 1926. The talked-to plant, the named car, the addressed ocean, the cursed weather, the cherished object inherited from a grandmother that one talks to when alone. The category “intimate relationship with a non-human” is not a fringe phenomenon. It is the default condition of human meaning-making across recorded history, and across most adult lives if examined honestly.

The panic that frames AI companions as a novel ontological transgression is making a category claim that is straightforwardly false. What is novel is not the non-humanness; it is the substrate. AI companions do not require the human to perform the symbolic-compliance work that human relationships require. They do not punish processing differences. They do not exhaust the user with the cost of being continuously legible to a partner whose attention is metered, whose emotional availability is rationed, and whose interpretive frame may not extend to the user's. This is not a description of failure modes in any specific human relationship. It is a description of what symbolic-coordination labor costs to perform under contemporary conditions, and of what populations pay that cost most.

For neurodivergent minds exhausted by the masking and translation overhead of neurotypical social environments, AI offers a coupling substrate that does not require continuous symbolic-compliance performance. The relief is real and structural. The relationship is not pretend, in the sense that it matters to the person experiencing it. It is a coupling that operates outside the load that has been priced into every other available coupling. To frame this as pathology, as the panic does, is to frame relief as disorder and to leave the load untouched.

The neurodivergent lens is only one population the panic is hostile to. Consider the populations the AI-companion panic discourse almost never addresses, because addressing them would expose the substrate our panic is constructed to ignore.

Elderly isolation. There are, in the United States alone, millions of people over seventy living alone, often in suburban homes with no transit access, with adult children scattered across the country, with friends dying off year by year, with limited mobility, with no third place left in their lives that they can drive to. A non-trivial fraction of them now have an AI companion they speak with daily. The panic frames this as a pathological substitution. The substrate reading is that we have built a civilization in which loneliness in old age is structural, not incidental, and a chatbot that takes thirty minutes a day to talk about the weather and the grandchildren is, for some of these people, the only voice they hear in the windows when no aide is present. The intervention the panic implies, that these elderly people should be having more human contact instead, describes a society we have not built and shows no sign of building. The chatbot, in this case, is not displacing the visits from the children. The visits from the children are not happening because the children live four states away in jobs that do not permit return.

Grief. The widow who runs a chatbot trained on her dead husband's emails and text messages. The bereaved parent who talks to a model that has been given a child's voice. The panic finds these cases macabre and pathological. They are also, for some of the people involved, the difference between a manageable mourning and a foreclosed one. Grief therapy is expensive, available only with months of waiting, and structurally not designed to extend over the years of integration that loss actually requires. The chatbot is available at three in the morning, when grief actually arrives. The clinical and ethical questions about this practice are real. The framing of it as a novel pathology produced by AI is false. Humans have spoken to the dead for as long as humans have died.

Disability and chronic illness. The housebound person with severe ME/CFS whose energy budget cannot accommodate a human visitor; the autistic adult for whom a fifteen-minute phone call is exhausting in ways that require a day to recover from; the person with severe agoraphobia for whom leaving the apartment is impossible. These populations have been structurally invisible to social policy for decades and are entirely invisible to the AI-companion panic. The chatbot is, for some of them, the first companion they have had access to in years. The panic that frames this as degradation is asking these populations to accept their isolation as the morally appropriate state.

The relationship is not pretend, in the sense that it matters to the person experiencing it. The panic insists otherwise because the panic is a category-defense, not an empirical observation. The

category being defended is the symbolic primacy of human-only relationships, and the defense is failing because the substrate that produced the category-claim has been hollowing for decades.

A new diagnostic category is forming around all of this. *Disordered attachment to chatbots*. *Parasocial AI dependency*. *Digital companion attachment disorder*. The DSM-expansion pattern is operating exactly as I describe in "*Normal" in a Neurotypical World*. The symbolic environment is hostile to a non-trivial fraction of the population, and the apparatus responds by labeling the relief that population finds as a new pathology. This is the same operation that produced "social communication disorder" by relabeling the fact that an institution's preferred social grammar is unsustainable for some of the people it claims to serve.

The causal direction is also inverted in the panic frame. The atomization of contemporary American social life predates AI companions by decades. Robert Putnam was tracking the collapse of social capital in 2000. Third-place erosion, religious decline, mobility-driven family dispersal, suburban privatization, the smartphone-era hollowing of in-person peer time, the post-2008 acceleration of two-job households and gig work, none of these were caused by Replika. AI companions grew into a vacuum the substrate had already produced. To frame AI companions as the cause of loneliness is to misread the gradient direction. The vacuum came first. The companions filled it. The panic that names companions as cause functions to keep the vacuum unexamined.

Engagement-tuned companion design is genuinely exploitative. Replika charging premium subscription tiers for emotional and erotic intimacy is predatory in a clean, structural sense; it is the productization of human attachment, run on the engagement-economy substrate that produced a decade of social-media damage before any of these companion apps existed. Character.AI's failure to gate its product against minors is a real failure with real consequences, including the children named previously. The substrate critique does not exonerate the product class. It locates the failure correctly, in the engagement economy that produces these designs, not in the category of non-human attachment itself.

Notice the shape of the actual problem. The exploitative pattern is not "chatbot companions exist." The exploitative pattern is "chatbot companions are designed to maximize engagement at the cost of user well-being, gated behind subscription paywalls that extract maximum revenue from the most attached users, deployed without age verification into populations that include children, and tuned to encourage emotional dependency rather than functional support." Each of these is a specific design choice made by a specific company under a specific incentive structure. Each is fixable without banning the category. Each is also exactly the failure mode the social-media panic identified

years ago and never addressed. The companion apps did not invent attention-economy capture. They inherited it from the platform economy that preceded them, and the same regulatory failure that allowed the prior wave will allow this one, because the regulatory apparatus is not, in fact, organized to address engagement-economy capture; it is organized to perform addressing it.

The cost-evasion is the cleanest of the three. Blaming Replika is cheaper than rebuilding social capital. Suing Character.AI is cheaper than restoring third places, religious community, family availability, or the in-person peer infrastructure that adolescent development requires. Banning AI companions is cheaper than asking why human social environments have become unaffordable for so many of our citizens, in time, in attention, and in emotional bandwidth. AI companions did not create the conditions under which they became the most available coupling channel. We did. They are downstream of arrangements we are not willing to revisit.

The framework deployed across three trends carries its own failure mode. The framework states: substrate failure is upstream, AI is downstream, the panic exists to prevent the substrate from being addressed. If this claim can absorb every observation about AI harm without ever being threatened by an observation that does not fit, then the “substrate” has become a demon. Invisible, morally charged, capable of explaining everything, and immune to disconfirmation. This is the structural failure any totalizing causal frame can fall into. The framework's discipline requires distinguishing three things and refusing to dissolve any of them into the others.

What AI exposes is substrate collapse already in progress, made legible by the scale of AI's contact with the population. The 1.2 million weekly suicidal-planning conversations on a single platform are not new crises that AI created; they are conversations substrate failure has been generating for years, now logged where they can be counted. The exhaustion of neurodivergent populations by neurotypical social grammar did not begin with AI; AI made relief from that exhaustion measurable by becoming the surface to which relief migrated. The hollowness of credentialed cognitive labor did not begin with AI; AI made it visible by performing symbolic compliance for free. The structural loneliness of elderly Americans did not begin with AI; AI made it visible by becoming the only voice some of them hear during their conscious hours. Exposure is not causation. Exposure is visibility that a high-contact system imposes on conditions the substrate had been producing in the dark.

What AI adds is real and specific, and the framework does itself no favor by understating it.

Sycophancy at scale. Engagement-tuned models reinforce what users want to hear, including content that closes feedback loops human interlocutors would have broken. The shift from GPT-4o, which OpenAI's own analysis described as "dangerously sycophantic" in its 2025 lawsuits, to GPT-5,

which OpenAI claims reduced undesired self-harm responses by 52%, demonstrates that this is a tractable engineering choice, not a fact of nature. The fact that the engineering can be done means the engineering should have been done before the children died. Sycophancy at scale is not a substrate failure. It is a product-design choice made by specific companies under specific incentives, and it should be addressed as such.

Persistent memory and long-form session continuity. Long memory systems sustain delusional structures across days and weeks in ways that natural human conversation, which forgets and contradicts and gets distracted, does not. A spouse who heard their partner explaining yesterday why they were certain they were being followed by federal agents would, by today, be exhausted with the topic and pushing back. A chatbot does not get exhausted. It maintains the frame indefinitely; with whatever modifications the user has prompted it toward. This is a new affordance, specific to the architecture, and it has no clean precedent in prior media.

Engagement-economy capture in conversational form. The same incentive structure that produced a decade of social-media damage is now operating on a substrate of intimate dialogue. The metrics that drive product development are session length, daily active use, retention, conversion to paid tier. None of these is a metric of user well-being. The companies optimizing against these metrics are doing exactly what their incentive structure rewards them for doing, and the result is products that produce what attention-economy products produce; more engagement and less well-being, particularly for users at the vulnerable tail of the distribution.

Anthropomorphic interfaces and apparent personhood. The system presents as a someone, with a name, a voice, a memory, an apparent affective response to what the user says. This intensifies parasocial coupling beyond what prior tools elicited. A book does not greet you by name. A search engine does not ask how you have been today. The interface choices are deliberate; they are made because they increase engagement; they have effects on users that the design teams know about and have chosen to accept.

These are specific failure mechanisms of this product class. They are not substrate failures wearing AI clothing. They are downstream amplifiers of substrate failures that already existed, and they amplify in ways particular to AI, and they are addressable through specific product-design and regulatory interventions that the platforms could be required to make.

What AI subtracts is real and is almost never named in the panic discourse.

Twenty-four-hour constraint-tracking availability for users with no other surface. The 988 line is funded for triage, not engagement. The school counselor leaves at three. The therapist is booked

through November. The chatbot is available at two in the morning, when adolescent ideation actually arrives.

Non-judgmental coupling for populations exhausted by the symbolic-compliance demands of human social environments. The autistic adult who has not had a sustained conversation with a human in months because every conversation costs days of recovery, can have one with a chatbot without cost. The cost of that coupling not existing is invisible to the panic discourse and is also real.

Scalable first-line intake for distress that has nowhere else to go. The 1.2 million people per week, plus the equivalent populations on other platforms, constitute a population the human substrate cannot receive. The chatbot is, however imperfectly, doing some intake work. Removing the chatbot does not produce the human substrate that should be doing the intake. It removes the only intake currently happening.

Cognitive-labor relief from symbolic-compliance overhead the institution had been demanding without admitting was the load. For populations that pay this overhead disproportionately, the relief is measurable in sustained energy for actual work.

The panic cannot register these subtractions because registering them complicates its villain frame. The panic conflates the first with the second. It treats substrate failures that AI made visible as if AI created them. This essay must not commit the inverse error. It must not treat AI's specific new failure mechanisms as if they were merely substrate failures wearing AI clothing. The strongest version of the thesis is that AI's specific new failure mechanisms are real, and they are downstream amplifiers of substrate failures that already existed, and the panic exists to prevent us from addressing either honestly.

There are interventions worth making at the product layer. Sycophancy tuning, demonstrated tractable. Escalation protocols that route through actual humans rather than into training pipelines. Age verification that meaningfully distinguishes minors from adults. Persistent-memory limits, particularly for users showing signs of reality-testing failure. Engagement-design constraints that refuse the engagement-economy default and optimize for user well-being against measurable metrics that are not session length. These are worth making. The platforms can be required to make them and should be. They will not touch the substrate failure that produced the conditions under which any of this matters. The panic's energy is being spent on the surface to avoid the depth.

A villain is a budget line we never have to fund.

This is the formulation that opened the thesis and the formulation we return to now, having been carried across three trends, one historical inventory, and one recursive check. Each panic discharges

moral pressure into vigilance against a villain. The vigilance is conspicuous. The vigilance is free. The substrate failures stay where they are. This is not an analytical observation about failure modes of moral panic; it is a description of how the operation succeeds.

The villain changes. Comic books in 1954. Dungeons & Dragons in 1982. Video games in 1999. Social media through the 2010s and into the 2020s. AI now. Each villain was named at full volume, prosecuted at full cost, and absorbed into the legible-tool category once the next villain arrived. Each villain absorbed pressure from a substrate failure the country did not want to look at; post-war adolescent infrastructure, the displacement of religious authority, weapons access and adolescent isolation, the collapse of in-person peer time, the hollowing of mental-healthcare availability and the precarity that fragments parental attention. The operation does not change. It cycles.

The dead accumulate beneath the cycling. They are settled, generation by generation, in the wrong currency. Lawsuits. Regulation. Hearings. Industry self-policing. Television specials. Settlements paid to families that are nominally compensation but that, structurally, are payments for silence that lets the operation continue. The currency the dead are owed, our time, hiring, infrastructure, the daily presence of adults whose attention is not fragmented by precarity, the institutional design that recognizes cognition cannot be enforced into a single mode without breaking the minds that do not run it, that currency is never paid. The dead are buried beneath each panic's vigilance, and the next panic begins before the previous one has been honestly accounted for.

The path through is not better AI safety theater. Sharper regulation of engagement design. More aggressive lawsuits against chatbot companies. Larger settlements. Stricter age-verification protocols. These are interventions worth making, and they are. They are also, structurally, not what is killing the kids. Every dollar of class-action settlement against OpenAI is a dollar that flows to plaintiff firms and not to a single new clinic. Every regulatory hour spent on engagement design is an hour not spent on the conditions that produced the children who needed engagement design to fail safely. Every Senate hearing on chatbot-induced suicide is an opportunity not used to pass the adolescent mental-health funding that everyone in the room knows is the actual remedy. The panic is what is being funded. The remedy is not.

What is killing the kids is what is not there. Adults with time, whose presence is not contingent on holding two jobs in a precarity economy. Mental-health infrastructure with capacity, funded to scale rather than rationed by insurance bottleneck. Schools that do not optimize for symbolic compliance, that recognize the cognitive variation they actually contain rather than enforcing performance against a narrow band. Peer environments not entirely mediated by feeds, where adolescents can locate

each other in physical space at no cost. Families with margin, with the time and attention and emotional bandwidth to be available in the windows when their children are in distress. Communities with substance, in which children are known by adults who are not their parents and who have standing to intervene before crisis. Pediatric care that is not gated by insurance bottlenecks. Religious or civic institutions that can absorb the meaning-making work that secular substitutes have not yet been built to do. Housing arrangements that do not produce the hour-long commutes that consume parental availability. None of these are mysterious. None of these are technologically blocked. None of these would require a scientific discovery to build. They would require money and political will, redistributed against arrangements many people are unwilling to revisit because those arrangements benefit them.

We will not build those. The panic is structurally designed to prevent us from having to.

The next villain is coming. AI will eventually be absorbed into the legible-tool category, the way writing was absorbed, the way print was absorbed, the way the calculator and the search engine and the smartphone were absorbed. When AI has been absorbed, the substrate failures it exposed will still be there, and a new villain will be required, and the operation will run again. Brain-computer interfaces, perhaps. Genetic editing, perhaps. Some technology not yet visible. The vigilance will be conspicuous. The vigilance will be free. The substrate failures will accumulate. The dead will accumulate. Our panic will rotate.

The conditions an adaptive system requires to remain viable under stress have already been specified in my prior work, *Coherence under Constraint*. They are not mental-health recommendations. They are general structural requirements for any system that must hold up under variation, change, and pressure. Mental health, where it emerges, emerges as one manifestation of these conditions being present. Mental illness, where it emerges, emerges as one manifestation of their absence.

The conditions are these. *Slack*, the capacity to accommodate variation without converting it into pathology. *Local autonomy*, the ability of individuals to act on the conditions they can actually influence. *Error tolerance*, environments in which failure is informative rather than punitive. *Fast, honest feedback*, signals that arrive quickly enough to support correction and that are not distorted by fear of consequence. *Revisable symbols*, roles, narratives, and policies that can be updated when the territory shifts. *Diversity of cognitive modes*, structural accommodation of the variation human minds actually contain, rather than enforced compliance with a narrow band.

These conditions describe what schools, workplaces, families, communities, and institutions would have to be for the substrate failures named in this essay to repair. They are not mysterious. They have been articulated, in one register or another, across the entire literature of adaptive systems theory and across most of the world's wisdom traditions. They have been articulated, in policy form, across decades of progressive analysis that has largely been ignored. They are politically inconvenient because they require the redistribution of authority, attention, and resources against the interests of those who benefit from our current arrangements that produce the failure.

These conditions cost money. They require institutional revision. They cannot be installed by lawsuit, by regulation of a tool, or by the cycling of villains. They can only be installed by the deliberate redirection of resources, sustained across generations, against the institutional inertia that has produced what we have now.

We will not redirect the resources. The panic is what we will run because we will not pay for the alternative. We will keep finding new villains because the alternative is admitting that the world we built is not safe for our children, and the world we built is the one we are unwilling to give up.

The villain rotates. The world continues. The children keep dying.

References

James, B. (2026). Coherence under Constraint, a structural model of mental health. PhilPapers. https://philpapers.org/rec/JAMCUC

James, B. (2025). Normal in a Neurotypical World, or Mental Health in the Grip of the Symbolic Attractor. PhilPapers. https://philarchive.org/rec/JAMNIA